

08-23-00

A

08/22/00  
 3621 U.S. PRO

PATENT

Attorney's Docket No.: U 012911-3

## IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Box Patent Application  
 Assistant Commissioner for Patents  
 Washington, D.C. 20231

3621 U.S. PRO  
 09/643407  
 08/22/00

## NEW APPLICATION TRANSMITTAL

Transmitted herewith for filing is the patent application of Inventors:

1. ITZHAK PEER
2. RON SHAMIR

**WARNING:** *The Declaration must name all of the actual inventor(s).*

For (title):

METHOD FOR SEQUENCING POLYNUCLEOTIDES

## 1. Type of Application

This new application is for a(n) (check one applicable item below):

- ☒ Original (nonprovisional)  
☐ Design  
☐ Plant

**WARNING:** *Do not use this transmittal for a completion in the U.S. of an International Application under 35 U.S.C. 371(c)(4) unless the International Application is being filed as a divisional, continuation or continuation-in-part application.***WARNING:** *Do not use this transmittal for the filing of a provisional application.*

## CERTIFICATION UNDER 37 CFR 1.10

I hereby certify that this New Application Transmittal and the documents referred to as enclosed therein are being deposited with the United States Postal Service on this date AUGUST 22, 2000 in an envelope as "Express Mail Post Office to Addressee" Mailing Label Number EL699731101US addressed to the: Assistant Commissioner of Patents, Washington, D.C. 20231

JENNIFER RAHSKIN

(type or print name of person mailing paper)

*Jennifer Resch*  
 (Signature of person mailing paper)

**NOTE:** *Each paper or fee referred to as enclosed herein has the number of the "Express Mail" mailing label placed thereon prior to mailing. 37 CFR 1.10(b).***WARNING:** *Certificate of mailing (first class) or facsimile transmission procedures of 37 CFR 1.8 cannot be used to obtain a date of mailing or transmission for this correspondence.*

(Application Transmittal [4-1]—page 1 of 7)

EXPRESS MAIL LABEL  
 NO.: EL699731101US

2. **Benefit of Prior U.S. Application(s) (35 U.S.C. 119(e), 120, or 121)**

**NOTE:** If the new application being transmitted is a divisional, continuation or a continuation-in-part of a parent case, or where the parent case is an International Application which designated the U.S., or benefit of a prior provisional application is claimed, then check the following item and complete and attach **ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED**.

**WARNING:** If an application claims the benefit of the filing date of an earlier filed application under 35 U.S.C. 120, 121 or 365(c), the 20-year term of that application will be based upon the filing date of the earliest U.S. application that the application makes reference to under 35 U.S.C. 120, 121 or 365(c). (35 U.S.C. 154(a)(2) does not take into account, for the determination of the patent term, any application on which priority is claimed under 35 U.S.C. 119, 365(a) or 365(b).) For a c-i-p application, applicant should review whether any claim in the patent that will issue is supported by an earlier application and, if not, the applicant should consider canceling the reference to the earlier filed application. The term of a patent is not based on a claim-by-claim approach. See Notice of April 14, 1995, 60 Fed. Reg. 20,195, at 20,205.

**WARNING:** When the last day of pendency of a provisional application falls on a Saturday, Sunday, or Federal holiday within the District of Columbia, any nonprovisional application claiming benefit of the provisional **must** be filed prior to the Saturday, Sunday or Federal holiday within the District of Columbia. See 37 C.F.R. § 1.78(a)(3).

- ☐ The new application being transmitted claims the benefit of prior U.S. application(s) and enclosed are **ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED**.

**NOTE:** If one of the following 3 items apply, then complete and attach **ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF A PRIOR U.S. APPLICATION CLAIMED** and a **NOTIFICATION IN PARENT APPLICATION OF THE FILING OF THIS CONTINUATION APPLICATION**.

- ☐ Divisional.  
☐ Continuation.  
☐ Continuation-in-Part (C-I-P).

3. **Papers Enclosed That Are Required For Filing Date Under 37 CFR 1.53 (Regular) or 37 CFR 1.153 (Design) Application**

24 Pages of specification

6 Pages of claims

1 Pages of Abstract

— Sheets of drawing

- ☐ formal  
☐ informal

**WARNING:** **DO NOT** submit original drawings. A high quality copy of the drawings should be supplied when filing a patent application. The drawings that are submitted to the Office must be on strong, white, smooth, and non-shiny paper and meet the standards according to § 1.84. If corrections to the drawings are necessary, they should be made to the original drawing and a high-quality copy of the corrected original drawing then submitted to the Office. Only one copy is required or desired. Comments on proposed new 37 CFR 1.84. Notice of March 9, 1988 (1990 O.G. 57-62).

**NOTE:** "Identifying indicia, if provided, should include the application number or the title of the invention, inventor's name, docket number (if any), and the name and telephone number of a person to call if the Office is unable to match the drawings to the proper application. This information should be placed on the back of each sheet of drawing a minimum distance of 1.5 cm. (5/8 inch) down from the top of the page." 37 C.F.R. 1.84(c).

(complete the following, if applicable)

- ☐ The enclosed drawing(s) are photograph(s), and there is also attached a "PETITION TO ACCEPT PHOTOGRAPH(S) AS DRAWING(S)". 37 C.F.R. 1.84(b).

4. **Additional papers enclosed**

- ☐ Preliminary Amendment
- ☐ Information Disclosure Statement (37 CFR 1.98)
- ☐ Form PTO-1449
- ☐ Citations
- ☐ Declaration of Biological Deposit
- ☐ Submission of "Sequence Listing," computer readable copy and/or amendment pertaining thereto for biotechnology invention containing nucleotide and/or amino acid sequence.
- ☐ Authorization of Attorney(s) to Accept and Follow Instructions from Representative
- ☐ Special Comments
- ☐ Other

5. **Declaration or oath**

- ☐ Enclosed

executed by (check **all** applicable boxes)

- ☐ inventors.
- ☐ legal representative of inventors. 37 CFR 1.42 or 1.43
- ☐ joint inventor or person showing a proprietary interest on behalf of inventor who refused to sign or cannot be reached.
  - ☐ This is the petition required by 37 CFR 1.47 and the statement required by 37 CFR 1.47 is also attached. See *item 13 below for fee*.

- ☒ Not Enclosed.

**WARNING:** Where the filing is a completion in the U.S. of an International Application but where a declaration is not available or where the completion of the U.S. application contains subject matter in addition to the International Application the application may be treated as a continuation or continuation-in-part, as the case may be, utilizing ADDED PAGE FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION CLAIMED.

- ☒ Application is made by a person authorized under 37 CFR 1.41(c) on behalf of **all the above named inventors**. (The declaration or oath, along with the surcharge required by 37 CFR 1.16(e) can be filed subsequently).

**NOTE:** It is important that all the correct inventor(s) are named for filing under 37 CFR 1.41(c) and 1.53(b).

- ☐ Showing that the filing is authorized. (Not required unless called into question. 37 CFR 1.41(d).)

6. **Inventorship Statement**

**WARNING:** If the named inventors are each not the inventors of all the claims an explanation, including the ownership of the various claims at the time the last claimed invention was made, should be submitted.

The inventorship for all the claims in this application are:

- ☐ The same
- ☐ Not the same. An explanation, including the ownership of the various claims at the time the last claimed invention was made,

7. **Language**

**NOTE:** An application including a signed oath or declaration may be filed in a language other than English. A verified English translation of the non-English language application and the processing fee of \$130.00 required by 37 CFR

1.17(k) is required to be filed with the application or within such time as may be set by the Office. 37 CFR 1.52(d).

NOTE: A non-English oath or declaration in the form provided or approved by the PTO need not be translated. 37 CFR 1.69(b).

- ☒ English
- ☐ non-English
- ☐ the attached translation is a verified translation. 37 CFR 1.52(d).

## 8. Assignment

☒ An assignment of the invention to RAMOT UNIVERSITY AUTHORITY FOR APPLIED RESEARCH &

- ☐ is attached. A separate ☐ "COVER SHEET FOR ASSIGNMENT (DOCUMENT) ACCOMPANYING NEW PATENT APPLICATION" or ☐ FORM PTO 1595 is also attached.

☒ will follow.

NOTE: "If an assignment is submitted with a new application, send two separate letters—one for the application and one for the assignment." Notice of May 4, 1990 (1114 O.G. 77-78).

**WARNING:** A newly executed "CERTIFICATE UNDER 37 CFR 3.73(b)" must be filed when a continuation-in-part application is filed by an assignee. Notice of April 30, 1993. 1150 O.G. 62-64.

## 9. Certified Copy

Certified copy of application

Country

Appln. No.

Filed

from which priority is claimed

- ☐ is attached.
- ☐ will follow.

NOTE: The foreign application forming the basis for the claim for priority must be referred to in the oath or declaration. 37 CFR 1.55(a) and 1.63.

NOTE: This item is for any foreign priority for which the application being filed directly relates. If any parent U.S. application or International Application from which this application claims benefit under 35 U.S.C. 120 is itself entitled to priority from a prior foreign application then complete item 18 on the ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED.

## 10. Fee Calculation (37 CFR 1.16)

A. ☒ Regular Application

---

Claims as Filed

---

Number Filed		Number Extra		Rate	Basic Fee 37 CFR 1.16(a) \$690.00
Total Claims (37 CFR 1.16(c))	32 - 20 =	12 x \$		18.00	216.00
Independent Claims (37 CFR 1.16(b))	3 - 3 =	1 x \$		<del>78.00</del>	<del>78.00</del>
Multiple dependent claim(s), if any (37 CFR 1.16(d))		+ \$		260.00	

- ☐ Amendment cancelling extra claims enclosed.
- ☐ Amendment deleting multiple-dependencies enclosed.
- ☐ Fee for extra claims is not being paid at this time.

**NOTE:** If the fees for extra claims are not paid on filing they must be paid or the claims cancelled by amendment, prior to the expiration of the time period set for response by the Patent and Trademark Office in any notice of fee deficiency. 37 CFR 1.16(d).

Filing Fee Calculation \$

- B. ☐ Design application  
(\$310.00 — 37 CFR 1.16(f))

Filing Fee Calculation \$

- C. ☐ Plant application  
(\$480.00 — 37 CFR 1.16(g))

Filing Fee Calculation \$

**11. Small Entity Statement(s)**

- ☐ Verified Statement(s) that this is a filing by a small entity under 37 CFR 1.9 and 1.27 is(are) attached or has been filed.

Filing Fee Calculation (50% of A, B or C above) \$

**NOTE:** Any excess of the full fee paid will be refunded if a verified statement and a refund request are filed within 2 months of the date of timely payment of a full fee. 37 CFR 1.28(a).

**12. Request for International-Type Search (37 CFR 1.104(d)) (Complete, if applicable)**

- ☐ Please prepare an international-type search report for this application at the time when national examination on the merits takes place.

**13. Fee Payment Being Made At This Time**

- ☒ Not Enclosed
- ☒ No filing fee is to be paid at this time. (This and the surcharge required by 37 CFR 1.16(e) can be paid subsequently.)

- ☐ Enclosed

☐ basic filing fee \$

- ☐ Recording assignment  
(\$40.00; 37 CFR 1.21(h)) (See attached "COVER SHEET FOR ASSIGNMENT ACCOMPANYING NEW APPLICATION.")
- ☐ Petition fee for filing by other than all the inventors or person on behalf of the inventor where inventor refused to sign or cannot be reached.  
(\$130.00; 37 CFR 1.47 and 1.17(h)) \$
- ☐ For processing an application with a specification in a non-English language.  
(\$130.00; 37 CFR 1.52(d) and 1.17(k)) \$
- ☐ Processing and retention fee  
(\$130.00; 37 CFR 1.53(d) and 1.21(l))
- ☐ Fee for international-type search report  
(\$40.00; 37 CFR 1.21(e)). \$

**NOTE:** 37 CFR 1.21(l) establishes a fee for processing and retaining any application which is abandoned for failing to complete the application pursuant to 37 CFR 1.53(d) and this, as well as the changes to 37 CFR 1.53 and 1.78, indicate that in order to obtain the benefit of a prior U.S. application, either the basic filing fee must be paid or the processing and retention fee of \$1.21(l) must be paid within 1 year from notification under §53(d).

Total fees enclosed \$

#### 14. Method of Payment of Fees

- ☐ Check in the amount of \$
- ☐ Charge Account No. 12-0425 in the amount of \$
- A duplicate of this transmittal is attached.

**NOTE:** Fees should be itemized in such a manner that it is clear for which purpose the fees are paid. 37 CFR 1.22(b).

#### 15. Authorization to Charge Additional Fees

**WARNING:** If no fees are to be paid on filing, the following items should not be completed.

**WARNING:** Accurately count claims, especially multiple dependent claims, to avoid unexpected high charges, if extra claim charges are authorized.

- ☐ The Commissioner is hereby authorized to charge the following additional fees by this paper and during the entire pendency of this application to Account No. 12-0425.
- ☐ 37 CFR 1.16(a), (f) or (g) (filing fees)
- ☐ 37 CFR 1.16(b), (c) and (d) (presentation of extra claims)

**NOTE:** Because additional fees for excess or multiple dependent claims not paid on filing or on later presentation must only be paid or these claims cancelled by amendment prior to the expiration of the time period set for response by the PTO in any notice of fee deficiency (37 CFR 1.16(d)), it might be best not to authorize the PTO to charge additional claim fees, except possibly when dealing with amendments after final action.

- ☐ 37 CFR 1.16(e) (surcharge for filing the basic filing fee and/or declaration on a date later than the filing date of the application)
- ☐ 37 CFR 1.17 (application processing fees)

**WARNING:** While 37 CFR 1.17(a), (b), (c) and (d) deal with extensions of time under §1.136(a), this authorization should be made only with the knowledge that: "Submission of the appropriate extension fee under 37 C.F.R. 1.136(a) is to no avail unless a request or petition for extension is filed." (Emphasis added). Notice of November 5, 1985 (1060 O.G. 27)

- ☐ 37 CFR 1.18 (issue fee at or before mailing of Notice of Allowance, pursuant to 37 CFR 1.311(b))

*NOTE: Where an authorization to charge the issue fee to a deposit account has been filed before the mailing of a Notice of Allowance, the issue fee will be automatically charged to the deposit account at the time of mailing the notice of allowance. 37 CFR 1.311(b).*

*NOTE: 37 CFR 1.28(b) requires "Notification of any change in loss of entitlement to small entity status must be filed in the application ... prior to paying, or at the time of paying, ... issue fee". From the wording of 37 CFR 1.28(b): (a) notification of change of status must be made even if the fee is paid as "other than a small entity" and (b) no notification is required if the change is to another small entity.*

**16. Instructions As To Overpayment**

- ☐ credit Account No. 12-0425  
☐ refund



Signature of Attorney

Reg. No. 25,858

Tel. No. (212) 708-1930

William R. Evans  
Ladas & Parry  
26 West 61 Street  
New York, NY 10023

☐ **Incorporation by reference of added pages**

*(Check the following item if the application in this transmittal claims the benefit of prior U.S. application(s) (including an international application entering the U.S. stage as a continuation, divisional or C-I-P application) and complete and attach the ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED)*

- ☐ Plus Added Pages for New Application Transmittal Where Benefit of Prior U.S. Application(s) Claimed

Number of pages added \_\_\_\_

- ☐ Plus Added Pages for Papers Referred to in Item 4 Above

Number of pages added \_\_\_\_

- ☐ Plus "Assignment Cover Letter Accompanying New Application"

Number of pages added \_\_\_\_

☒ **Statement Where No Further Pages Added**

*(If no further pages form a part of this Transmittal, then end this Transmittal with this page and check the following item:)*

- ☒ This transmittal ends with this page.

**PATENT**

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re application: ITZHAK PEER, et al

For: METHOD FOR SEQUENCING POLYNUCLEOTIDES

Attorney Docket No.: U 012911-3

**Assistant Commissioner for Patents  
Washington, D.C. 20231**

Sir:

**PRELIMINARY AMENDMENT**

Please amend the above application as follows:

**IN THE CLAIMS**

Claim 3, line 1, delete "or 2"

Claim 6, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 7, line 1, delete "any one of Claims 1 to 6" and replace therefor -- claim

1--

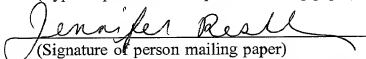
---

**CERTIFICATE UNDER 37 1.10**

I hereby certify that this paper is being deposited with the United States Postal Service on this date AUGUST 22, 2000 in an envelope as "EXPRESS MAIL POST OFFICE TO ADDRESS-EE" Mailing Label Number EL699731101US addressed to the: Commissioner of Patents and Trademarks, Washington, D.C. 20231

JENNIFER RASHKIN

(Type or print name of person mailing paper)

  
(Signature of person mailing paper)

**NOTE:** Each paper or fee referred to as enclosed herein has the number of the "EXPRESS MAIL" mailing label place thereon prior to mailing 37 CFR 1.16(b).

EXPRESS MAIL LABEL

NO.: EL699731101US

Claim 8, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 13, line 1, delete "any one of Claims 1 to 7" and replace therefor

-- claim 1--

Claim 15, line 1, delete "any one of Claims 1 to 7" and replace therefor

-- claim 1--

Claim 19, line 1, delete "any one of Claims 1 to 11" and replace therefor

-- claim 1--

Claim 20, line 1, delete "any of Claims 13, 14, or 19" and replace therefor

-- claim 13--

Claim 21, line 1, delete "any of Claims 15, 17, or 18" and replace therefor

-- claim 15--

Claim 22, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 23, line 1, delete "any one of Claims 1 to 22" and replace therefor

-- claim 1--

Claim 24, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 25, line 1, delete "any one of Claims 1 to 5, 8 to 24" and replace

therefor -- claim 1--

Claim 26, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 27, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 28, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 29, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Claim 30, line 1, delete "any one of the previous claims" and replace therefor

-- claim 1--

Respectfully submitted,

  
\_\_\_\_\_  
WILLIAM R. EVANS  
LADAS & PARRY  
26 WEST 61<sup>ST</sup> STREET  
NEW YORK, NEW YORK 10023  
REG.NO.25858(212)708-1930

## METHOD FOR SEQUENCING POLYNUCLEOTIDES

## FIELD OF THE INVENTION

This invention relates to computational methods in molecular biology, and more specifically to methods for determining the sequence of a polynucleotide.

## 5 REFERENCES

- Baines, W., and Smith, G.C., *J. Theor. Biology*, **135**:303-307 (1988).
- Ben-Dor, A., Pe'er, I., Shamir, R., and Sharan, R., In *Proceedings of the Tenth International Conference on Combinatorial Pattern Matching (CPM '99)*, 88-100  
10 (1999), New York: ACM Press.
- Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q., and Lander, E.S., *Nature Genetics*,  
15 **22**:231-238 (1999).
- Drmanac, R., and Crkvenjakov, R., Yugoslav Patent Application 570 (1987).
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of proteins and Nucleic Acids*, Cambridge University Press,  
20 (1998).
- Eddy, S.R., *Current Opinions in Structural Biology*, **6**(3):361-365 (1996).
- 25 Hirschberg, D.S., *Communications of the ACM*, **18**,6:341-343 (1975).
- Jukes, T.H., and Cantor, C.R., *Mammalian Protein Metabolism*, New York: Academic Press 21-123 (1969).
- 30 Khrapko, K.R., Lysov, Y.P., Khorlyn, A.A., Shick, V.V., Florentiev, V.L., and Mirzabekov, A.D., *FEBS Letters*, **256**:118-122 (1989).
- Kimura, M., *Journal of Molecular Evolution*, **16**:111-120 (1980).

Krogh, A., Brown, M. Mian, S., Sjolander, M., and Haussler, D., Applications to protein modeling. Technical Report UCSC-CRL-93-32, Department of Computer and Information Sciences, University of California at Santa Cruz (1993).

- 5 Krogh, A., Brown, M., Mian S., Sjolander, M. and Haussler, D., Appliations to protein modeling **235(5)**:1501-1531 (1994).

Lysov, Y., Floretiev, V., Khorlyn, A., Khrapko, K., Shick, V., and Mirzabekov, A., *Dokl. Acad. Sci.*, USSR, **303**:1508-1511 (1988).

10

Macevices, S.C., International Patent Application PS US89 04741 (1989).

National Center for Biotechnology Information, 2000, A database of single nucleotide polymorphisms, <http://www.ncbi.nlm.nih.gov/SNP/>

15

Pevzner, P.A., and Lipshutz, R.J., *Mathematical Foundations of Computer Science*, LNCS **841**:143-158 (1994).

20

Pevzner, P.A., Lysov, Y. P., Khrapko, K.R., Belyavsky, A.V., Florentiev, V.L., and Mirzabekov, A.D., *J. Biomol. Struct. Dyn.* **7**:63-73 (1989).

Preparata, F., Frieze, A., and Upfal, E., *Journal of Computational Biology* **6(3-4)**:361-368 (1999).

25

Skiena, S.S., and Sundaram, G., *J. Comput. Biol.* **2**:333-353 (1995).

Smith, T.F., and Waterman, M.S., *Journal of Molecular Biology* **147(1)**:195-197 (1981).

30

Southern, E.M., Maskos, U., and Elder, J.K., *Genomics* **13**:1008-1017 (1992).

Southern E., UK patent Application GB 8,810,400 (1988).

Southern, E.M., *Trends in Genetics* **12**:110-115 (1996).

35

Wang, D.G., Fan, J., Siao, C., Berno, A., Young P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Lipshutz, R., Chee, M., and Lander, E.S., *Science* **280**:1077-1082.

40

Yang, Z., *Molecular Biology and Evolution* **10**:1396-1401.

## BACKGROUND OF THE INVENTION

Sequencing by hybridization (SBH) is a method for sequencing a polynucleotide such as a DNA molecule (Bains & Smith 1988, Lysov *et al.* 1988, Southern 1988, Drmanac and Crkvenjakov 1987, Macevics 1989). In this method, a chip, or microarray is used consisting of a surface upon which all possible oligonucleotide probes of a particular length  $k$  (referred to herein as " $k$ -mers") are immobilized (Southern 1996). The DNA molecule whose sequence is to be determined, referred to as the "*target molecule*", is allowed to hybridize to the  $k$ -mers on the chip. The target molecule and the  $k$ -mers on the chip may all be single stranded molecules. Alternatively, a double stranded target may first be cut into fragments having single stranded "*sticky ends*", and the  $k$ -mers on the chip may be the sticky ends of double stranded molecules. Ideally, a single stranded target or the sticky end of a double stranded target hybridizes to a  $k$ -mer on the chip if and only if the sequence complementary to the  $k$ -mer occurs somewhere in the target sequence or the sticky end. Thus, in principle, it is possible to experimentally determine the "*k-spectrum*" of the target (the set of all  $k$ -long substrings present in the target). In practice, however, the data are ambiguous due to the ability of the target to bind to  $k$ -mers that are only partially complementary to one of its substrings. Thus, any binarization of the hybridization signal will contain errors.

The goal of SBH is to determine the target sequence from the target spectrum. However, even if the target spectrum were error free, the target sequence is not uniquely determined by the spectrum. If the number of sequences consistent with the spectrum is large, there is no satisfactory method to select the true sequence. Theoretical analysis and simulations (Southern *et al.*, 1992, Pevzner and Lipshutz 1994) have shown that even when the spectrum is errorless and the correct multiplicity of each  $k$ -mer in the target sequence is known, the average length of a uniquely reconstructible target sequence using a chip of 8-mers is only about two hundred nucleotides, far below the length of a DNA molecule that may be sequenced by electrophoresis.

Let  $\Sigma = (A, C, G, T)$  designate the set of nucleotides composing a DNA molecule.  $M = 4$  is the "alphabet size". A DNA sequence is a string over  $\Sigma$  which is denoted herein between braces ( $\langle \rangle$ ). The  $k$ -spectrum of a target sequence  $T$  of length  $L$ ,  $T = \langle t_1, t_2, \dots, t_L \rangle$ , is the set of all  $k$ -long substrings ( $k$ -mers) of  $T$ . For each  $k$ -mer  $\vec{x} = \langle x_1, x_2, \dots, x_k \rangle$  in  $\Sigma^k$ , we define  $T(\vec{x})$  to be 1 if  $\vec{x}$  is a substring of  $T$ , and 0 otherwise. We denote  $K = M^k$ , the number of  $k$ -mers. A hybridization experiment measures, for each  $k$ -mer  $\vec{x}$  in  $\Sigma^k$ , an intensity of its hybridization with the target.

The result of an SBH experiment may be described by a graph in which each candidate target sequence is represented as a path in a graph (Pevzner *et al.*, 1989). The graph is a directed de-Bruijn graph  $G(V, E)$  whose vertices are labeled by all the  $(k-1)$ -mers (the set of vertices  $V = \Sigma^{k-1}$ ), and its edges are labeled by  $k$ -mers, (the set of edges  $E = \Sigma^k$ ). The edge labeled  $\langle x_1, x_2, \dots, x_k \rangle$  connects the vertex  $\langle x_1, x_2, \dots, x_{k-1} \rangle$  to the vertex  $\langle x_2, x_3, \dots, x_k \rangle$ . There is a 1-1 correspondence between  $L$ -long candidate target sequences and  $(L - k + 1)$ -long paths in  $G$ , whose edge labels comprise the target spectrum. Hereafter, we interchangeably refer to edges and their labels, and also to sequences and their corresponding paths.

Since  $k$ -mers may reoccur in the target sequence, the paths do not have to be simple. When the spectrum is perfect and the multiplicities of the  $k$ -mers in the spectrum are known, every solution is an Eulerian path (Pevzner *et al.* 1989). In practice, however, the spectrum is not perfect and the multiplicities are not known.

Alternative chip designs (Bains and Smith 1988, Khrapko *et al.* 1989, Pevzner *et al.* 1991, Preparata *et al.* 1999, Ben-Dor *et al.* 1999), as well as interactive protocols (Skiena and Sundaram 1995) have been suggested, often assuming additional information, in order to reduce the ambiguity of the hybridization-based reconstruction.

Nucleotide sequences from different sources may resemble each other, due to a common ancestral gene. This phenomenon is encountered within a species,

between duplicated regions within a genome, and between individuals within a population. Small differences in sequences, referred to as "*Single Nucleotide Polymorphisms*" or *SNPs*, efficiently serve as genetic markers that are useful in medicine. Thus the detection and genotyping of SNPs has become an important task of human geneticists. The evolution of homologous sequences from a common ancestral gene is mainly due to nucleotide substitution. Insertions and deletions of nucleotides are also known to have occurred during evolution of homologous sequences, though at lower rates.

A DNA molecule having a known sequence and known to be homologous to a target molecule has not yet been used to reduce the ambiguity of SBH data in order to determine the target sequence.

## SUMMARY OF THE INVENTION

In the following description and set of claims, two parameters are considered to be equivalent to each other if they are proportional to each other.

The present invention provides a method for sequencing a target sequence. In accordance with the invention, experimental spectrum data obtained from a DNA chip is combined with sequence information of a reference DNA molecule. The reference molecule is preferably a molecule believed to be homologous with the target. For example, the target sequence may be a mutant gene and the reference sequence the previously sequenced normal gene. As another example, the target sequence may be a human gene and the reference sequence the homologous gene in another organism. A score is defined for each sequence in a set of candidate target sequences based upon a simultaneous comparison of the candidate sequence with the spectrum and with the reference sequence. A candidate target sequence is then selected having a essentially maximal score. Calculating the score does not require knowledge of the multiplicities of the k-mers in the k-spectrum. Moreover, unlike all prior art algorithms, the score does not assume that the spectrum is perfect.

The invention therefore provides a novel probabilistic method that handles imperfect hybridization data with unknown multiplicities. Thus, in accordance with the invention the hybridization of the target T with the k-mer on the DNA chip complementary to  $\bar{x}$  is described by probabilities  $P_0(\bar{x})$  and  $P_1(\bar{x})$  of the observed hybridization signal when  $T(\bar{x}) = 0$ , and  $T(\bar{x}) = 1$ , respectively. The results of a hybridization experiment are described by the "probabilistic spectrum" (PS) defined as the pair  $(P_0, P_1)$  of functions  $P_i: \Sigma^k \rightarrow [0, 1]$ . If the experiment were perfect, i.e., if  $P_0(\bar{x})$  and  $P_1(\bar{x})$  are either 0 or 1 with  $P_0(\bar{x}) + P_1(\bar{x}) = 1$ , then the PS would represent the k-spectrum. In practice, however,  $P_0(\bar{x})$  and  $P_1(\bar{x})$  are both positive. There is thus a chance  $1 - P_0(\bar{x})$  for a false positive (a k-mer  $(\bar{x})$  not occurring in T, whose complementary sequence produces a hybridization signal indicative of hybridization) and a chance  $1 - P_1(\bar{x})$  for a false negative (a k-mer  $(\bar{x})$  occurring in T, whose complementary sequence produces a signal indicative of no hybridization). (When handling probabilities, some of which are perfect, problems of division by zero might occur. This is avoided by implicitly perturbing probabilities 0 and 1 to  $\epsilon$  and  $1 - \epsilon$ .)

The probability of obtaining a specific spectrum PS when T is used as the target is referred to as the "experimental likelihood". The experimental likelihood is calculated assuming that the hybridization results of the target to different k-mer probes are mutually independent. In one embodiment of the invention, an experimental likelihood  $L^e(\hat{T})$  is used that does not assume knowledge of the multiplicities of each k-mer in the sequence.  $L^e(\hat{T})$  is given by:

$$L^e(\hat{T}) = \text{Prob}(\text{PS} | \hat{T}) = \prod_{\bar{x} \in \Sigma^k} P_{\hat{T}(\bar{x})}(\bar{x}) \quad (1)$$

Taking logarithms and defining  $\omega(\bar{x}) = \log \frac{P_1(\bar{x})}{P_0(\bar{x})}$  we can write:

$$\log P_{\hat{T}(\bar{x})}(\bar{x}) = \begin{cases} \log P_0(\bar{x}) & \text{if } \hat{T}(\bar{x}) = 0 \\ \log P_0(\bar{x}) + \omega(\bar{x}) & \text{if } \hat{T}(\bar{x}) = 1. \end{cases} \quad (2a)$$

Hence,

$$\log L^e(\hat{T}) = \sum_{\bar{x} \in \sum^k} \log P_0(\bar{x}) + \sum_{\hat{T}(\bar{x})=1} \omega(\bar{x}). \quad (2b)$$

5 The first term is a constant (independent of  $\hat{T}$ ), and is omitted hereafter.

In another embodiment, an approximate likelihood  $\tilde{L}(\hat{T})$  is used, that is defined as follows: Let  $p = e_0, \dots, e_{L-k}$  be the path in  $G$  corresponding to  $\hat{T}$  and define

$$\log \tilde{L}^e(\hat{T}) = \sum_{i=0}^{L-k} \omega(e_i). \quad (3)$$

10  $\tilde{L}^e(\hat{T}) = L^e(\hat{T})$  for a path in which all edges have a multiplicity of 1, and is otherwise an approximation to  $L^e(\hat{T})$ .  $\tilde{L}^e(\hat{T})$  has the advantage of being easily computable in a recursive manner:

$$\log \tilde{L}^e(e_0, \dots, e_l) = \log \tilde{L}^e(e_0, \dots, e_{l-1}) + \omega(e_l) \quad (4)$$

15 In yet another embodiment, an experimental likelihood  $\underline{L}^e(\hat{T})$  is used that takes into account the multiplicities of edges. In this case, the probabilistic spectrum consists of probabilities  $P_i(\bar{x})$ , denoting the probability of the observed hybridization signal when the multiplicity of  $\bar{x}$  in the target is  $i$ .  $\underline{L}^e(\hat{T})$  is defined by:

$$\underline{L}^e(\hat{T}) = Prob(PS | \hat{T}) = \prod_{\bar{x} \in \sum^k} P_{\hat{T}(\bar{x})}(\bar{x}) \quad (4b)$$

20 where  $\hat{T}(\bar{x})$  is the multiplicity of  $\bar{x}$  in  $\hat{T}$ .

Thus in its first aspect, the invention provides a method for obtaining a candidate sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule T, T producing a hybridization signal  $I(\bar{x})$  upon incubating T with a polynucleotide  $\bar{x}$  for each polynucleotide  $\bar{x}$  in a set E of polynucleotides, the method comprising the steps of:

(a) for each polynucleotide  $\bar{x}$  in the set E of polynucleotides, obtaining a probability  $P_0(\bar{x})$  of the hybridization signal  $I(\bar{x})$  when the sequence  $\bar{x}$  is not complementary to a subsequence of T and a probability  $P_1(\bar{x})$  of the hybridization signal when the sequence  $\bar{x}$  is complementary to a subsequence of T; so as to obtain a probabilistic spectrum (PS) of T;

(b) assigning a score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) selecting one or more candidate nucleotide sequences having an essentially maximal score.

In its second aspect, the invention provides a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for obtaining a candidate nucleotide sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule T, T producing a hybridization signal  $I(\bar{x})$  upon incubating T with a polynucleotide  $\bar{x}$  for each polynucleotide  $\bar{x}$  in a set E of polynucleotides, the method comprising the steps of:

(a) for each polynucleotide  $\bar{x}$  in the set E of polynucleotides, obtaining a probability  $P_0(\bar{x})$  of  $I(\bar{x})$  when the sequence  $\bar{x}$  is not complementary to a subsequence of T and a probability  $P_1(\bar{x})$  of  $I(\bar{x})$  when the sequence  $\bar{x}$  is complementary to a subsequence of T; so as to obtain a probabilistic spectrum (PS) of T;

(b) assigning a score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) selecting a candidate nucleotide sequence having an essentially maximal score.

In its third aspect the invention provides a computer program product comprising a computer useable medium having computer readable program code embodied therein for obtaining a candidate nucleotide sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule T, T producing a hybridization signal  $I(\bar{x})$  upon incubating T with a polynucleotide  $\bar{x}$  for each polynucleotide  $\bar{x}$  in a set E of polynucleotides, the computer program product comprising:

10 (a) for each polynucleotide  $\bar{x}$  in the set E of polynucleotides, computer readable program code for causing the computer to obtain a probability  $P_0(\bar{x})$  of  $I(\bar{x})$  the sequence  $\bar{x}$  is not complementary to a subsequence of T and a probability  $P_1(\bar{x})$  of  $I(\bar{x})$  when the sequence  $\bar{x}$  is complementary to a subsequence of T;

(b) computer readable program code for causing the computer to assign a  
15 score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) computer readable program code for causing the computer to select a candidate nucleotide sequence having an essentially maximal score.

20

## DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

### First Embodiment

In this embodiment, the unknown target sequence  $T = \langle t_1 \dots t_i \rangle$  has a known, homologous reference sequence  $H = \langle h_1 \dots h_i \rangle$ . H and T are known to  
25 differ from each other by nucleotide substitutions without insertions or deletions (indels). This would be the case, for instance, when the target T is a mutant sequence whose wild type sequence H has already been sequenced, and one expects that nucleotide substitutions are the only cause of variability between H and T (statistically, substitutions are much more prevalent than indels (Wang *et*

al. 1998)). A set of  $M \times M$  position specific substitution matrices  $M^{(1)}, \dots, M^{(l)}$  are used, where for each position  $j$  along the sequence:

$$M^{(j)}[i, i'] = Prob(t_j = i \mid h_j = i') \quad (5)$$

for nucleotides  $i$  and  $i' \in \Sigma$ .

The matrices  $M^{(j)}$  may be the same for all  $j$ , or may differ for different positions  $j$ . The matrices  $M^{(j)}$  are used to calculate a distribution on the space of possible target sequences. This "prior distribution for ungapped homology",  $D^u$ , is given, for each candidate target sequence  $T$  by:

$$D^u(\hat{T}) = Prob(\hat{T} \mid H) = \prod_{j=1}^l M^{(j)}[t_j, h_j] \quad (6)$$

One may recursively compute:

$$D^u(\langle t_1 \dots t_j \rangle) = \langle t_1 \dots t_{j-1} \rangle \cdot M^{(j)}[t_j, h_j] \quad (7)$$

We denote  $L^{(j)}[x, y] \equiv \log M^{(j)}[x, y]$ .

The probability of a candidate target sequence  $\hat{T}$ , given the probability spectrum  $PS$  and the reference sequence  $H$  is:

$$Prob(\hat{T} \mid H, PS) = \frac{Prob(H) \cdot Prob(\hat{T} \mid H) \cdot Prob(PS \mid H, \hat{T})}{Prob(H, PS)} \quad (8)$$

Given  $\hat{T}$ , the hybridization signal is independent of  $H$ :

$$Prob(PS \mid H, \hat{T}) = Prob(PS \mid \hat{T})$$

Thus, omitting the constant  $\frac{Prob(H)}{Prob(H,PS)}$  we can write:

$$Prob(\hat{T} | H, PS) \cong D^u(\hat{T}) \cdot L^e(\hat{T}) \quad (9a)$$

$$Prob(\hat{T} | H, PS) \cong D^u(\hat{T}) \cdot \tilde{L}^e(\hat{T}) \quad (9b)$$

$$5 \quad \text{or} \quad Prob(\hat{T} | H, PS) \cong D^u(\hat{T}) \cdot \underline{L}^e(\hat{T}) \quad (9c)$$

Taking logarithms, the following "*ungapped scores*" of a candidate target are obtained:

$$Score_1^u(\hat{T}) = \log L^e(\hat{T}) + \log D^u(\hat{T}) \quad (10a)$$

$$10 \quad Score_2^u(\hat{T}) = \log \tilde{L}^e(\hat{T}) + \log D^u(\hat{T}) \quad (10b)$$

$$Score_3^u(\hat{T}) = \log \underline{L}^e(\hat{T}) + \log D^u(\hat{T}) \quad (10c)$$

With  $Score_1^u$ ,  $Score_2^u$  or  $Score_3^u$ , the higher the score of a sequence  $\hat{T}$ , the more likely it is to be the target sequence. The highest scoring candidate sequence may be determined by any method known in the art. In the search for the highest scoring candidate sequence, complexity is preferably reduced by deleting from the graph edges for which  $\tilde{L}^e(\hat{T})$ ,  $L^e(\hat{T})$  or  $\underline{L}^e(\hat{T})$  is less than a predetermined constant. Isolated vertices corresponding to highly improbable (k-1)-mers, are also preferably deleted from the graph.

20 For example, using  $\tilde{L}^e(\hat{T})$ , the search for a high scoring candidate sequence may be performed by the following algorithm referred to herein as "*Algorithm A*". In accordance with Algorithm A, for each vertex  $\bar{y} = \langle y_1 \dots y_{k-1} \rangle \in \Sigma^{k-1}$ , and integer  $j = k - l, k, k + l, \dots, l$ , let  $S^u[\bar{y}, j]$  be the maximum score of a  $j$ -long sequence ending with  $\bar{y}$  aligned to  $\langle h_1 \dots h_l \rangle$ . Initialize, for each  $\bar{y}$ :

$$S^u[\bar{y}, k-1] = \sum_{j=1}^{k-1} L^{(j)}[y_j, h_j] \quad (11)$$

Loop over  $j = k, \dots, l$ , and for each vertex  $\bar{y} = \langle y_1 \dots y_{k-1} \rangle$  recursively update:

$$S^u[\bar{y}, j] = L^{(j)}[y_{k-1}, h_j] + \max_{e=(\bar{z}, \bar{y}) \in E} \{S^u[\bar{z}, j-1] + \omega(e)\} \quad (12a)$$

Finally, return:

$$MAX\ Score^u = \max_{\bar{y} \in V'} S^u[\bar{y}, l] \quad (12b)$$

10

A sequence  $T^*$  attaining the maximal score is found from the matrix  $S^u$  as is known in the art, for example, by saving trace-back pointers:

$$P[\bar{y}, j] = \arg \max_{\bar{z} = \langle z_1 \dots z_{k-1} \rangle, \bar{z} \in E(\bar{z}, \bar{y})} \{S^u[\bar{z}, j-1] + \omega(e)\} \quad (13a)$$

$$MAXPtr = \arg \max_{\bar{y} \in V'} S^u[\bar{y}, l] \quad (13b)$$

15

The maximum-scoring path in the graph is then followed, by setting:

$\bar{z}^j = MAXPtr$ , and for all  $j = k, \dots, l$  :  $\bar{z}^{j-1} = P[\bar{z}^j, j]$ . Denote  $\bar{z}^j = \langle z_1^j \dots z_{k-1}^j \rangle$ .

The final result is the sequence of nucleotides  $\langle z_1^{k-l}, z_2^{k-l}, \dots, z_{k-1}^{k-l}, z_{k-l}^k, z_{k-l-1}^{k-l}, \dots, z_{k-l}^l \rangle$

20

The time complexity is  $O(lK)$ , since the maximization in (12a), (13a) is a maximum of only a constant number (four) of terms. Although the complexity is exponential in  $k$ , it is constant for a given microarray (currently feasible values are  $k = 8$  or  $9$ ). Moreover, the complexity scales linearly with the size of the hybridization experimental results, which are part of the input.

Space complexity requires a more elaborate analysis. When naively using this algorithm, it requires  $O(lK)$  memory space, which is quite high for current technology microarrays. We now detail how we can modify the algorithm to reduce space complexity.

- 5 Observe, that this algorithm consists of two computations: Computing the optimal score (equations (11),(12a) and (12b)), and reconstructing the optimal sequence (equations (13a) and (13b)). The first task, of computing the optimal score alone, is space-efficient: it can be accomplished using space which is linear in the (effective) size of the hybridization experimental data, that is,  $O(K)$
- 10 space.

By following the paradigm of Hirschberg (Hirschberg 1975), for example, for linear-space pair-wise alignment, a version of the algorithm is obtained which requires only linear space. The reduced space complexity is traded for time complexity, which increases by an  $O(\log l)$  factor.

- 15 For each position  $j = l, l-1, \dots, k, k-1$ , the score of the entire sequence is decomposed. The total score is represented as a sum of two expressions: the contribution of its  $(j - k + 1)$ -prefix, which equals the score of this prefix computed by  $S^u$ , plus the contribution of the corresponding suffix. Formally, for each vertex  $\bar{y} = \langle y_1 \dots y_{l-1} \rangle \in V$ , let  $R^u[\bar{y}, j]$  be the maximum contribution to the
- 20 score of a  $(l - j + k - 1)$ -long sequence beginning with  $\bar{y}$  aligned to  $\langle h_{l-k+2} \dots h_l \rangle$ . Initialize, for each  $\bar{y}$ :

$$R^u[\bar{y}, l] = 0 \quad (14)$$

- 25 Loop over  $j = l - 1, l - 2, \dots, k - 1$ , and for each vertex  $\bar{y} = \langle y_1 \dots y_{l-1} \rangle$  recursively update:

$$R^u[\bar{y}, j] = \max_{e = (\bar{y}, \bar{z}) \in E} \left\{ R^u[\bar{z}, j+1] + \omega(e) + L^{(j+1)}[z_{k-1}, h_{j+1}] \right\} \quad (15)$$

Observe that, for all  $k - l \leq j \leq l$

$$MAX Score^u = \max_{\vec{y} \in V} \{S^u[\vec{y}, j] + R^u[\vec{y}, j]\} \quad (16)$$

5

Equation (16) can be used to decompose the problem into two similar problems, of half its size. Recursively solving these sub-problems gives a divide-and-conquer approach for finding the optimal sequence. The linear space algorithm is therefore as follows:

- 10      1. If the length  $l$  of the target is smaller than some constant  $C$ , for example, 25 nucleotides:  
Solve the problem directly, according to the dynamic program of Equations (11), (12a), (12b), (13a) and (13b).
- Otherwise,
- 15      2. Set  $m = \frac{l+k-1}{2}$ .
3. For each  $j = k - l, \dots, m$ :  
Compute  $S^u[\vec{y}, j]$  (following equations (11) and (12a)) for all  $\vec{y}$ , re-using space.
4. For each  $j = l, l - l, \dots, m$ :  
20      Compute  $R^u[\vec{y}, j]$  (following equations (14) and (15)) for all  $\vec{y}$ , re-using space.
5. Find  $\vec{y}_m = \arg \max_{\vec{y} \in V} \{S^u[\vec{y}, m] + R^u[\vec{y}, m]\}$ ,  
thereby computing:  $MAX Score^u$ , by (16).
6. Recursively compute:
  - 25      (a) The optimal sequence aligned to  $\langle h_1 \dots h_m \rangle$   
ending with  $\vec{y}_m$ .
  - (b) The optimal sequence aligned to  $\langle h_m \dots h_l \rangle$

beginning with  $\bar{y}_m$ .

Observe, that for each  $\bar{y}, j$ , the values of  $S^u[\bar{y}, j]$  and  $R^u[\bar{y}, j]$  are computed a total of  $\log l$  times. Thus the algorithm takes  $O(K/l \log l)$  time and  $O(K)$  space, using the effective spectrum.

### Second Embodiment: substitutions and deletions

In this embodiment, the unknown target sequence  $T = \langle t_1 \dots t_r \rangle$  differs from the reference  $H = \langle h_1 \dots h_l \rangle$ , by substitutions and deletions only, without insertions.

Denote the probability of initiating a gap right before  $h_j$  (aligning  $h_j$  to space) is  $2^{\alpha_j}$ . Similarly,  $\beta_j$  is the logarithm of the probability for gap extension at  $h_j$ . Also define  $\hat{\beta}_j = \log(1 - 2^{\beta_j})$ ,  $\hat{\alpha}_j = \log(1 - 2^{\alpha_j})$ . To overcome boundary problems at the ends of the sequence, we extend the alphabet by including left and right space characters:  $\bar{\Sigma} = \Sigma \cup \{\triangleright, \triangleleft\}$ . We augment the reference sequence by the string  $\triangleright^k$  on its left and  $\triangleleft^k$  on the right. We extend the substitution matrix by using probabilities that force alignment of each of  $\triangleright$  and  $\triangleleft$  to itself. Formally, we define:

$$\begin{aligned} \overline{\Sigma^{k-1}} &= \Sigma^{k-1} \cup \left\{ \bar{x}\bar{z} \mid \bar{x} = \triangleright^j, \bar{z} \in \Sigma^{k-1-j} \right\} \\ &\cup \left\{ \bar{z}\bar{x} \mid \bar{z} \in \Sigma^j, \bar{x} = \triangleleft^{k-1-j} \right\} \end{aligned} \quad (17)$$

20

We arbitrarily set  $\omega(\bar{y})$  to 0 for each  $\bar{y} \in \overline{\Sigma^{k-1}} \setminus \Sigma^{k-1}$ . Thus, the weighted de-Bruijn graph is naturally extended over  $\overline{\Sigma^{k-1}}$ , and so is  $[G] = ([V], [E])$ , its effective subgraph. Hereafter, we use the notation  $[G]$  for the extended graph. As with the previous embodiment, in order to reduce complexity, edges for which  $\tilde{L}^e(\hat{T})$  or  $L^e(\hat{T})$  is less than  $\varepsilon$  are preferably deleted from the graph. Isolated

vertices corresponding to highly improbable (k-1)-mers, are also preferably deleted from the graph.

The search for a high scoring candidate sequence may be performed by the following algorithm referred to herein as "*Algorithm B*". In accordance with  
 5 Algorithm B, for each  $\vec{y} = \langle y_1 \dots y_{k-l} \rangle \in [V]$ ,  $j = k = 1, k, k + 1, \dots, l$ ,  $S^d[\vec{y}, j]$  is defined as the maximum score of aligning a sequence ending with  $\vec{y}$  to  $\langle h_1 \dots h_j \rangle$  where  $h_j$  is aligned to a gap (and  $y_{k-1}$  is aligned to some  $h_1 \dots h_j$ ). Further  $T^d[\vec{x}, j]$  is defined as the maximum score of aligning a sequence ending with  $\langle y_1 \dots y_{k-1} \rangle$ , to  $\langle h_1 \dots h_j \rangle$  where  $h_j$  aligned to  $y_{k-l}$ . Initialize, for each  $\vec{y}$ :

10

$$S^d[\vec{y}, k-1] = -\infty; \quad (19)$$

$$T^d[\vec{y}, k-1] = \begin{cases} 0 & \vec{y} = \triangleright^{k-1} \\ -\infty & \text{otherwise} \end{cases} \quad (20)$$

15 Loop over  $j = k, \dots, l$ , and for each  $\vec{y} = \langle y_1 \dots y_{k-l} \rangle \in [V]$ , recursively update:

$$S^d[\vec{y}, j] = \max \{ T^d[\vec{y}, j-1] + \alpha_j, S^d[\vec{y}, j-1] + \beta_j \} \quad (21)$$

$$20 \quad T^d[\vec{y}, j] = L^{(j)}[y_{k-1}, h_j] + \max_{e = (\vec{z}, \vec{y}) \in E} \left\{ \omega(e) + \max \left\{ T^d[\vec{z}, j-1] + \hat{\alpha}_j, S^d[\vec{z}, j-1] + \hat{\beta}_j \right\} \right\} \quad (22)$$

Finally, return:

$$MAX Score^d = T^d \left[ \triangleleft^{k-1}, I \right] \quad (23)$$

The complexity of this algorithm is still  $O(I/K)$  and a linear space variant can be obtained, as described in the previous embodiment. A sequence  $T^*$  attaining the maximal score is then formed from the matrix  $T^d$  as is known in the art, for example, by saving trace-back pointers to follow the maximally scoring path in analogous manner to that described in the previous embodiment.

### Third Embodiment: Substitutions, Deletions and Insertions.

In this embodiment, a target sequence is determined when the target is known to be obtained from the reference by substitutions, insertions and deletions. The algorithm is an extension of the dynamic programs of the previous embodiments.

Denote by  $T_j$  the target prefix whose last nucleotide is aligned to  $h_j$  in the reference sequence. Further denote by  $a_j$  (respectively  $b_j$ ) the log-probability of initiating (extending) an insertion in the target after  $T_j$ , and define  $\hat{a}_j = 1 - a_j$ ,  $\hat{b}_j = 1 - b_j$ .

Consider the weighted graph  $(G, \omega)$ . Define the  $K \times K$  matrix  $W$  as follows:

$$W[\bar{x}, \bar{y}] = \begin{cases} 2^{w(\bar{y})} & \text{The } (k-1) - \text{suffix of } \bar{x} \\ & \text{is the } (k-1) - \text{prefix of } \bar{y} \\ 0 & \text{Otherwise} \end{cases} \quad (24)$$

$W^i[\bar{x}, \bar{y}]$  is thus the probability of moving from  $\bar{x}$  to  $\bar{y}$  along  $i$  edges. The probability of an insertion of length  $i$  after  $T_j$  is  $a_j b_j^i \hat{b}_j$ . Suppose that the prefix  $T_j$  ends with  $\bar{x}$ . Then  $a_j b_j^{i-1} \hat{b}_j W^i[\bar{x}, \bar{y}]$  is the probability of  $T_{j+1}$  ending with  $\bar{y}$

and being  $i$  nucleotides longer than  $T_j$ . The matrix  $W$  governing the stochastic progression from  $T_j$  to  $T_{j+1}$  is calculated as follows:

$$W' = \hat{a}_j b_j W^2 \hat{b}_j + a_j b_j^2 W^3 \hat{b}_j \dots \quad (25)$$

5

$$= \hat{a}_j W + a_j b_j \hat{b}_j W^2 \sum_{i \geq 2} b_j^{i-2} W^{i-2} \quad (26)$$

$$= \hat{a}_j W + a_j b_j \hat{b}_j W^2 (I - b_j W)^{-1} \quad (27)$$

10 A new weighted graph  $(G', \omega')$  is now defined as follows. The vertex set of  $G$  is also the vertex set of  $G'$ . The edge set  $E'$  of  $G'$  is the set of all pairs  $\bar{x}, \bar{y}$  with  $W'[\bar{x}, \bar{y}] > 0$ . Each such edge  $e = [\bar{x}, \bar{y}]$  is associated with a weight  $w'(e) = \log W'[\bar{x}, \bar{y}]$ .

The search for a high scoring candidate sequence may be performed by the  
15 following algorithm referred to herein as "*Algorithm C*". In accordance with Algorithm C, Algorithm B of the second embodiment is applied to  $(G', \omega')$  instead of  $(G, \omega)$ .

In contrast to  $G$ , degrees in  $G'$  are not bounded by 4. Therefore, computing each dynamic program cell has complexity  $O(K)$  in the worst case, with the total  
20 complexity of the algorithm being  $O(l|E'|)$ . Again, considering only the effective size of the graph allows more efficient computation, taking  $O(l|E'|)$ .

#### **Fourth Embodiment: Substitutions, Deletions and Insertions.**

In this embodiment, homology between nucleotide sequences is described by Hidden Markov Models (HMMs) using a set  $Q$  of Markov chain states with a  
25 predefined set of allowed transitions between them. For each level (position along the sequence)  $j = 1, \dots, l_Q$ ,  $Q$  includes three states:  $M_j$  (match),  $I_j$  (insert), and  $D_j$  (delete).  $M_j$  and  $D_j$  can be reached from the three  $(j-1)$  (th) level states.  $I_j$

can be reached from the three ( $j$ )-(th) level states (including a self-loop). Transition probabilities are as described in previous sections, e.g.,  $a_j = \text{Prob}(M_j \mapsto I_j)$ . Additionally, each insert or match state,  $q$ , induces a vector of emission probabilities  $M^q$ , where  $M^q[i]$  is the probability that the target nucleotide is  $i$ . We denote  $L^q[i] \equiv 0$  for  $q = D_j$ ,  $L^q[i] \equiv \log M^q[i]$  otherwise. We write  $lpb(X) \equiv \log \text{Prob}(X)$  for short.

The search for a high scoring candidate sequence may be performed by the following algorithm referred to herein as "*Algorithm D*". In accordance with Algorithm D, a three dimensional array  $S$  is defined, where for each  $q \in Q$ ,  $\bar{y} = \langle y_1 \dots y_{k-1} \rangle \in [V]^r$ ,  $r = k, \dots, L$ ,  $S[q, \bar{y}, r]$  is defined as the maximum score of an  $r$ -long sequence ending with  $\langle y_1 \dots y_{k-1} \rangle$ , whose alignment to the profile ends in  $q$ . Thus, initialize:

$$S[q_{start}, \triangleright^{k-1}, k-1] = 0 \quad (28)$$

$$S[q, \bar{y}, k-1] = -\infty \quad \text{for other values of } \bar{y}, q \quad (29)$$

Loop over  $r = k, \dots, L$ , and for each  $\bar{y} = \langle y_1 \dots y_{k-1} \rangle \in [V]^r$ ,  $r \leq L_Q$ , recursively update:

$$S[q, \bar{y}, r] = L^q[y_{k-1}] + \max_{\substack{e=(E,q) \in E \\ q'q \mapsto q}} \{S[q', \bar{z}, r-1] + lpb(q' \mapsto q) + \omega(e)\} \quad (30)$$

Finally, return:

$$MAX \text{ Score} = \max_l \{S[q_{end1}, \triangleleft^{k-1}, l]\} \quad (31)$$

A sequence  $T^*$  is maximal score is then found in a manner similar to that described in the previous embodiments.

This algorithm requires  $O(l_Q \cdot [K] \cdot L)$  time and space, where  $L$  is an upper bound on the size of the target sequence. As with the previous embodiments, the complexity of this algorithm can be reduced to  $O(l_Q \cdot [K] \cdot L \log L)$  time and  $O(l_Q \cdot [K])$  memory. Furthermore, one can consider the dynamic program as filling a  $l_Q \times L$  matrix, with a  $[K]$ -long vector in each matrix cell. Since all values far from the main diagonal of this matrix should be negligible, preferably only values within a distance less than a predetermined constant  $R$  to the main diagonal are calculated, reducing the complexity to  $O(R(l_Q+L) \cdot [K] \cdot \log L)$  time and  $O(R(l_Q+L) \cdot [K])$  space.

#### Fifth Embodiment: Summation over all paths

In this embodiment the graph nodes (HMM states and k-mers) that are most likely to be visited at a certain position along the target sequence are obtained. The “*Forward-Backward*” algorithm is used (see, e.g., Durbin et al., 1998) providing the likelihood summed over all paths entering a node, instead of the likelihood of the maximum path. The only change to the equation presented thus far is that *max* operators are changed into *log-sum-of-exponents*. More specifically, equations (12a), (12b), (15), (16), (20), (21), (29), and (30) are re-written, respectively, as follows:

$$S^u[\bar{y}, j] = L^{(j)}[y_{k-1}, h_j] + \log \sum_{e=(\bar{z}, \bar{y}) \in E} \exp(S^u[z, j-1] + \omega(e)) \quad (12a')$$

$$MAX\ Score^u = \log \sum_{\bar{y} \in V} \exp(S^u[\bar{y}, l]) \quad (12b')$$

$$R^u[\bar{y}, j] = \log \sum_{e=(\bar{y}, \bar{z}) \in E} \exp(R^u[\bar{z}, j+1] + \omega(e) + L^{(j+1)}[z_{k-1}, h_{j+1}]) \quad (15')$$

$$MAX\ Score^u = \log \sum_{\bar{y} \in V} \exp(S^u[\bar{y}, j] + R^u[\bar{y}, j]) \quad (16')$$

$$S^d[\bar{y}, j] = \log(\exp(T^d[\bar{y}, j-1] + \alpha_j) + \exp(S^d[\bar{y}, j-1] + \beta_j)) \quad (20')$$

$$T^d[\bar{y}, j] = L^{(j)}[y_{k-1}, h_j] + \log \sum_{e=(\bar{z}, \bar{y}) \in B} \exp(\alpha(e)) + \log(\exp(T^d[\bar{z}, j-1] + \hat{\alpha}_j) + \exp(S^d[\bar{z}, j-1] + \hat{\beta}_j)) \quad (21')$$

$$S[q, \bar{y}, r] = L^q[y_{k-1}] + \sum_{\substack{e=(E, q) \in E \\ q' \mapsto q}} \exp(S[q', \bar{z}, r-1] + lpb(q' \mapsto q) + \omega(e)) \quad (29')$$

$$MAX\ Score = \log \sum_i \exp(S[q_{end1} \prec^{k-1}, i]) \quad (30')$$

5

### Sixth Embodiment: Enhancements

In this embodiment the exact likelihood calculated according to Equation 10a of several top-scoring candidates found using the approximated likelihood (Equation 10b) is calculated. These sequences are then re-ranked. This 2-phase  
10 filtering is more discriminative than approximated scoring, while still tractable using the formulae presented.

If the score of a dynamic programming cell is very low, that cell probably does not participate in the maximum solution. This allows discarding such cells, without risking loss of the true optimum. Computing time and space may thus be  
15 saved.

The invention may be used for simultaneously re-sequencing several short targets, instead of a single long sequence. This scenario arises when considering many exons of a single gene. The invention may also be generalized to deal with DNA chips that do not contain the set of all k-mers.

20 When the set of oligonucleotides on the microarray is not the set of all k-mers, a graph is constructed consisting, as vertices, instead of all the (k-1)-mers, all the prefixes and suffixes of oligonucleotides on the microarray. Edges in this graph connect two vertices if there is one base pair suffix (suffix) addition to one of them, that makes the other its proper suffix (prefix). The

scoring mechanism remains the same. This also applies for oligonucleotides containing “gaps” or “universal bases” (Preparata et al., 1999).

The invention may be used also for sequencing polypeptides. Given a polypeptide chain homologous to a target, and given a collection of probabilities of occurrence of sub-chains along the target, our algorithms will find the optimal candidate target sequence.

### Example

The invention was implemented and tested on simulated data. Nucleotide substitutions were equiprobable and insertions and deletions were not allowed.

As a reference sequence, prefixes of the gene-rich human mitochondrial sequence, (Accession Number J01415) were used. For each reference sequence, the following were generated:

1. A target sequence generated according to a prescribed probability  $q$  of substitution, defining the matrix  $M$  as  $1-q$  on the diagonal and  $q/3$  elsewhere.
2. An 8-spectrum for the target was generated using the probabilistic spectrum defined by  $P_i(\bar{x}) = 1 - p$  if  $T(\bar{x}) = i$ , where  $p$  is a fixed probability.

All probabilistic parameters were constant, i.e., position/ $k$ -mer independent. For each 8-spectrum and target sequence, candidate sequences were scored using Eq. (10), and a candidate sequence of maximal score was found.

The algorithm was implemented in C++ and executed on Linux and SGI machines. Running times, on a Pentium 3, 600MHz machine, were roughly  $0.12 / \log l$  seconds for an  $l$ -long reference sequence (ranging from roughly 7 minutes for a 500bp-long sequence to 2.5 hours for 6Kb). Only the main memory was used, with the application consuming at most 40Mb. The graph was not reduced to its effective size. This would have reduced both space and time dramatically, at the expense of possibly missing the truly maximal scoring sequence.

The performance of the algorithm was quantified by the following figures of merit:

1. Full success rate-The fraction of runs for which the target sequence was perfectly reconstructed.
2.  $\epsilon$ -success rate - The fraction of runs for which the target sequence of length  $l$  was reconstructed with fewer than  $\epsilon \cdot l$  nucleotide errors.
3. Average sequencing error - The fraction of nucleotide errors.

Table 1 presents results for a scenario of distinct, but closely related sequences, e.g., orthologous genes in a pair of primates. We assume perfect hybridization data with 97% sequence similarity (that is  $q=0.03$ ). The results show that sequences of length up to 2000 can be reconstructed almost perfectly.

The non-monotonicity of the figures of merit with respect to the target length is probably due to sequence contents.

Table 2 presents results for a scenario of SNP-genotyping. The rate of SNPs is assumed to be 1:700 (Wang *et al.* 1998), and  $p=2\%$  was used. The results show that a high success rate is achievable even in the presence of spectrum errors.

**Table 1**

Length	# runs	% full success	% $\epsilon$ -success		% avg. error
			$\epsilon = 10^{-3}$	$\epsilon = 2 \cdot 10^{-3}$	
500	10	100	100	100	0.000
1000	10	100	100	100	0.000
1500	10	100	100	100	0.000
2000	17	94	94	94	0.003
2500	13	46	53	69	0.295
3000	14	71	78	78	0.488
3500	5	0	20	20	4.091
4000	13	76	84	84	2.173
4500	11	9	18	45	0.091
5000	15	0	13	53	4.149
5500	7	14	28	71	0.119

- Table 1: Results on simulated data, for different sequence lengths, assuming 97% sequence similarity between the target and the reference, and perfect hybridization data.

**Table 2**

Length	# runs	% full success	% $\epsilon$ -success		% avg. error
			$\epsilon = 10^{-3}$	$\epsilon = 2 \cdot 10^{-3}$	
250	10	100	100	100	0.000
500	10	100	100	100	0.000
750	10	90	90	100	0.013
1000	10	90	90	90	0.010
1250	10	90	100	100	0.032
1500	12	91	100	100	0.033
1750	10	60	80	80	0.109
2000	10	60	90	90	4.965
2500	10	0	80	100	10.312
3000	10	30	70	90	0.230

Table 2: Results on simulated data, for different sequence lengths, assuming  $p = 2\%$  error the hybridization data, with 1:700 sequence difference.

It will also be understood that the system according to the invention may be a suitably programmed computer. Likewise, the invention contemplates a computer program being readable by a computer for executing the method of the invention. The invention further contemplates a machine-readable memory tangibly embodying a program of instructions executable by the machine for executing the method of the invention.

**CLAIMS:**

1. A method for obtaining a candidate nucleotide sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule T, T producing a hybridization signal  $I(\bar{x})$  upon incubating T with a polynucleotide  $\bar{x}$  for each polynucleotide  $\bar{x}$  in a set E of polynucleotides, the method comprising the steps of:

(a) for each polynucleotide  $\bar{x}$  in the set E of polynucleotides, obtaining a probability  $P_0(\bar{x})$  of the hybridization signal  $I(\bar{x})$  when the sequence  $\bar{x}$  is not complementary to a subsequence of T and a probability  $P_1(\bar{x})$  of the hybridization signal when the sequence  $\bar{x}$  is complementary to a subsequence of T; so as to obtain a probabilistic spectrum (PS) of T;

(b) assigning a score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) selecting one or more candidate nucleotide sequences having an essentially maximal score.

2. The method according to Claim 1, wherein the polynucleotides  $\bar{x}$  in the set E are immobilized on a surface.

3. The method according to Claim 1 or 2, wherein the set E is a set of k-mers.

4. The method according to Claim 3 wherein E is the set of all k-mers formed from nucleotides from a predetermined set of nucleotides..

5. The method of Claim 4 wherein the predetermined set of nucleotides is selected from the group consisting of

(a) adenine, guanine, cytosine, and thymine; and

(b) adenine, guanine, cytosine, uracil.

6. The method according to any one of the previous claims, wherein the score of a candidate nucleotide sequence  $\hat{T}$  is based upon  $L^s(\hat{T})$  where

$$L^s(\hat{T}) = \prod_{\bar{x} \in \mathcal{A}} P_{\hat{T}(\bar{x})}(\bar{x}),$$

wherein  $\hat{T}(\bar{x})=0$  if the sequence of  $\bar{x}$  is not complementary to a subsequence of  $\hat{T}$  and  $\hat{T}(x)=1$  if the sequence of  $\bar{x}$  is complementary to a subsequence of  $\hat{T}$ .

7. The method according to any one of Claims 1 to 6, wherein the score of a candidate sequence  $\hat{T}$  is based upon  $\tilde{L}^e(\hat{T})$  where  $\log \tilde{L}^e(\hat{T}) = \sum_{i=0}^m \omega(e_i)$ , wherein  $\hat{T}$

5 contains polynucleotides  $e_0, \dots, e_m$  and  $\omega(e_i) = \log \frac{P_1(e_i)}{P_0(e_i)}$ .

8. The method according to any one of the previous claims, wherein T and H have a common length.

9. The method according to Claim 8, wherein the score of a candidate sequence  $\hat{T}$  is based upon  $D^u(\hat{T})$  where  $D^u(\hat{T}) = \prod_{j=1}^l M^{(j)}[t_j, h_j]$ , wherein  $M^{(j)}[t_j, h_j]$

10 is a probability of a nucleotide  $t_j$  in position j of T being replaced with nucleotide  $h_j$  in position j of H.

10. The method according to Claim 9, wherein the score of a candidate nucleotide sequence  $\hat{T}$  is  $\text{Score}_1^u(\hat{T})$ , or  $\text{Score}_2^u(\hat{T})$  where  $\text{Score}_1^u(\hat{T}) = \log L^e(\hat{T}) + \log D^u(\hat{T})$  and  $\text{Score}_2^u(\hat{T}) = \log \tilde{L}^e(\hat{T}) + \log D^u(\hat{T})$ .

15 11. The method according to Claim 10 wherein the polynucleotides in the set E are k-mers and the step of selecting a candidate nucleotide sequence having an essentially maximal score comprises the steps of

(a) For each (k-1)-mer  $\bar{y}$  calculating  $S^u[\bar{y}, k-1] = \sum_{j=1}^{k-1} L^{(j)}[y_j, h_j]$

(b) for each integer j = k, ..., l,

(ba) for each polynucleotide sequence  $(y_1, \dots, y_{k-1})$

(baa) calculating

$$S^u[\bar{y}, j] = L^{(j)}[y_{k-1}, h_j] + \max_{e=(\bar{z}, \bar{y}) \in E} \{S^u[\bar{z}, j-1] + \omega(e)\}$$

wherein  $L^{(j)}[y, h_j] = \log M^{(j)}[y, h_j]$ .

(bab) selecting a (k-1)-mer  $P[\bar{y}_j]$

satisfying

$$S^u[P[\bar{y}, j], j-1] + \alpha(P[\bar{y}, j], \bar{y}) = \max_{e=(\bar{z}, \bar{y}) \in E} \{S^u[z, j-1] + \alpha(e)\}$$

(c) selecting a  $(k-1)$ -mer  $Z^1$  having a score essentially equal to

$$\max_{j \in V} S^v[\bar{y}, l];$$

(d) for  $j=k-1, \dots, l-1$ ; recursively calculating  $(k-1)$ -mers  $Z^j$  where  $Z^{j-1} = P[Z^j, j]$

(e) selecting candidate target sequence  $\langle Z^{k-1}_1, Z^{k-1}_2, \dots, Z^{k-1}_{k-1}, Z^k_{k-1}, Z^{k+1}_{k-1}, \dots, Z^l_{k-1} \rangle$ , where  $Z^j = \langle Z^j_1, Z^j_2, \dots, Z^j_{k-1} \rangle$

**12.** The method according to Claim 9, wherein the polynucleotides in the set E are  $k$ -mers, and the step of selecting a candidate nucleotide sequence having an essentially maximal score comprises the steps of:

(a) If the length  $l$  of the target is greater than the predetermined constant, setting  $m = \frac{l+k-1}{2}$ ;

(b) For each  $j = k-l, \dots, m$ , computing  $S^u[\bar{y}, j]$  according to Claim 10 for all  $\bar{y}$ ;

(c) For each  $j = l, l-1, \dots, m$ , computing  $R^u[\bar{y}, j]$  according to equations (14) and (15) for all  $\bar{y}$ ;

(d) Selecting  $\bar{y}_m = \arg \max_{\bar{y} \in V} \{S^u[\bar{y}, m] + R^u[\bar{y}, m]\}$ ;

(e) Computing the optimal sequence aligned to  $\langle h_1 \dots h_m \rangle$  ending with  $\bar{y}_m$ , and the optimal sequence aligned to  $\langle h_1 \dots h_l \rangle$  beginning with  $\bar{y}_m$ .

**13.** The method according to any one of Claims 1 to 7, wherein H and T have lengths such that the length of T is less than the length of H.

**14.** The method according to Claim 13, wherein the step of assigning a score to each of a plurality of candidate nucleotide sequences and the step of selecting the candidate target sequence are performed according to Algorithm B.

15. The method according to any one of Claims 1 to 7, wherein H and T have arbitrary lengths.

16. The method according to Claim 15, wherein the step of assigning a score to each of a plurality of candidate nucleotide sequences and the step of selecting the candidate target sequence are performed according to Algorithm C.

17. The method according to Claim 15, wherein the step of assigning a score to each of a plurality of candidate nucleotide sequences and the step of selecting the candidate target sequence are performed according to Algorithm D.

18. The method according to Claim 17 wherein a Hidden Markov Model is used instead of a reference sequence.

19. The method according to any one of Claims 1 to 11, wherein the algebraic equation (12a') replaces the algebraic equation (12a), the algebraic equation (12b') replaces the algebraic equation (12b), the algebraic equation (15') replaces the algebraic equation (15), and the algebraic equation (16') replaces the algebraic equation (16).

20. The method according to any one of Claims 13,14, or 19, wherein the algebraic equation (20') replaces the algebraic equation (20), and the algebraic equation (21') replaces the algebraic equation (21).

21. The method according to any one of Claims 15, 17, or 18, wherein the algebraic equation (29') replaces the algebraic equation (29), and the algebraic equation (30') replaces the algebraic equation (30).

22. The method according to any one of the previous claims wherein the target comprises two or more polynucleotide molecules.

23. The method according to any one of Claims 1 to 22 computing the exact score  $L^e(\hat{T})$  for several candidate sequences chosen according to the value of the approximated score  $\tilde{L}^e(\hat{T})$ .

24. The method according to any one of the previous claims further comprising a step of deleting candidate sequences having likelihood below a predetermined value.

25. The method according to any one of Claims 1 to 5, 8 to 24, wherein the score of a candidate nucleotide sequence  $\hat{T}$  is based upon  $\underline{L}^e(\hat{T})$  where

$$\underline{L}^e(\hat{T}) = \prod_{\vec{x} \in A} P_{\hat{T}(\vec{x})}(\vec{x}),$$

wherein  $\hat{T}(\vec{x}) = r$  if the sequence of  $\vec{x}$  is complementary to exactly  $r$  subsequences

5 of  $\hat{T}$ .

26. The method according to any one of the previous claims, wherein the set  $E$  of polynucleotide does not include all the polynucleotide of a specific length.

27. The method according to any one of the previous claims, wherein the set  $E$  of polynucleotide includes polynucleotides of different lengths.

10 28. The method according to any one of the previous claims for use in a task selected from the group comprising:

- (a) Detecting or genotyping of Single Nucleotide Polymorphisms.
- (b) Detecting or genotyping of genetic syndroms or disorders.
- (c) Detecting or genotyping somatic mutations.

15 (d) Sequencing a polynucleotide whose function is related to the function of the reference polynucleotide.

29. The method according to any one of the previous claims, wherein polynucleotides contain gaps, or universal bases.

30. The method according to any one of the previous claims, wherein  
20 polypeptides are sequenced instead of polynucleotides.

31. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for obtaining a candidate nucleotide sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule  $T$ ,  $T$  producing a  
25 hybridization signal  $I(\vec{x})$  upon incubating  $T$  with a polynucleotide  $\vec{x}$  for each polynucleotide  $\vec{x}$  in a set  $E$  of polynucleotides, the method comprising the steps of:

- (a) for each polynucleotide  $\vec{x}$  in the set  $E$  of polynucleotides, obtaining a probability  $P_0(\vec{x})$  of  $I(\vec{x})$  when the sequence  $\vec{x}$  is not complementary to a subsequence of  $T$  and a probability  $P_1(\vec{x})$  of  $I(\vec{x})$  when the sequence  $\vec{x}$  is

complementary to a subsequence of T; so as to obtain a probabilistic spectrum (PS) of T;

(b) assigning a score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) selecting a candidate nucleotide sequence having an essentially maximal score.

**32.** A computer program product comprising a computer useable medium having computer readable program code embodied therein for obtaining a candidate nucleotide sequence, the candidate nucleotide sequence being indicative of a sequence of a target polynucleotide molecule T, T producing a hybridization signal  $I(\bar{x})$  upon incubating T with a polynucleotide  $\bar{x}$  for each polynucleotide  $\bar{x}$  in a set E of polynucleotides, the computer program product comprising:

(a) for each polynucleotide  $\bar{x}$  in the set E of polynucleotides, computer readable program code for causing the computer to obtain a probability  $P_0(\bar{x})$  of  $I(\bar{x})$  the sequence  $\bar{x}$  is not complementary to a subsequence of T and a probability  $P_1(\bar{x})$  of  $I(\bar{x})$  when the sequence  $\bar{x}$  is complementary to a subsequence of T;

(b) computer readable program code for causing the computer to assign a score to each of a plurality of candidate nucleotide sequences, the score being based upon the probabilistic spectrum and upon at least one reference nucleotide sequence H; and

(c) computer readable program code for causing the computer to select a candidate nucleotide sequence having an essentially maximal score.

**ABSTRACT**

A method for obtaining a nucleotide sequence that is indicative of the sequence of a target polynucleotide molecule T. The method makes use of hybridization data obtained by incubating T with nucleotide probes. A score is assigned to each of a plurality of candidate nucleotide sequences based upon the hybridization data and upon at least one reference nucleotide sequence. A candidate nucleotide sequence is then selected having an essentially maximal score.